Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease

Muhammad Minoar Hossain [a,*], Reshma Ahmed Swarna [a], Rafid Mostafiz [b], Pabon Shaha [a], Lubna Yasmin Pinky [a], Mohammad Motiur Rahman [a], Wahidur Rahman [c], Md. Selim Hossain [d], Md. Elias Hossain [e], Md. Sadiq Iqbal [f]

[a] *Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Bangladesh*
[b] *Department of Computer Science and Engineering, Dhaka International University, Bangladesh*
[c] *Department of Computer Science and Engineering, Khwaja Yunus Ali University, Bangladesh*
[d] *Department of Computing and Information System, Daffodil International University, Bangladesh*
[e] *Department of Software Engineering, Daffodil International University, Bangladesh*
[f] *Department of Computer Science and Engineering, Bangladesh University, Bangladesh*

ARTICLE INFO

Keywords:
Chronic kidney disease (CKD)
Feature optimization
Feature selection
Machine learning (ML)
Feature reduction
Feature importance

ABSTRACT

Chronic kidney disease (CKD) slowly decreases one's kidney ability. A machine learning (ML) based early CKD diagnosis scheme can be an effective solution to reduce this harm. The efficiency of ML techniques depends on the selection and use of the appropriate features. Hence, this research analysis several feature optimization approaches along with a max voting ensemble model to establish a highly accurate CKD diagnosis system by using an appropriate feature set. The ensemble model of this research is structured with five existing classifiers. Three types of feature optimization namely feature importance, feature reduction, and feature selection where for each approach two most proficient techniques are analyzed with the mentioned ensemble model. Based on all analysis the research gets a feature optimization technique called Linear discriminant analysis belonging to the feature selection approach provides the most outstanding result of 99.5% accuracy by using 10-fold cross-validation. The results of this research indicate the efficiency of feature optimization for the diagnosis of ML-based CKD.

1. Introduction

Chronic kidney disease (CKD) refers to the most dangerous step in the kidney damaging process. The human kidney gradually lost functionality and can be stopped working due to CKD. There are some risk factors for CKD like high blood pressure, cardiovascular illness, diabetes, age limit, and family history of kidney failure. On the other hand, obesity, autoimmune diseases, systemic infections, urinary tract infection, and kidney-related disorders like kidney loss, damage, injury, or infection are considered as the secondary risk factors to CKD. The treatment procedures also vary depending on the patient's physical condition. The fundamental treatments are changing the patient's lifestyle, medicine for controlling related problems, dialysis, and finally kidney transplant. According to a report of the year 2021, around 37 million population are estimated for having CKD only in the US (Chronic, 2021). It has been estimated that around 10% of people over the world are suffering from CKD (Almasoud & Ward, 2019). Around 2.4 million deaths occur every year due to kidney disease (Nikhila, 2021). Early diagnosis is a primary solution to provide treatment for CKD. Two common CKD diagnosis approaches are blood and urine tests. However, these tests are manual processes and require an expert to complete the process. Hence in recent years, various research has been going to develop the computerized automated CKD diagnosis approaches by utilizing the efficiency of artificial intelligence. In this regard Machine learning (ML) is a first choice by the researchers.

In recent years, several studies have been developed to establish ML-based CKD diagnosis. Hence, we have investigated several existing approaches for gaining appropriate guidance for this research. To establish a CKD diagnosis approach, Polat, Mehr, and Cetin (2017) analyzed the Support vector machine (SVM) classifier with two feature selection approaches namely filter and wrapper. Based on the evaluation the researchers found that SVM with filter method of best-first search

[*] Correspondence to: Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh.

E-mail addresses: minoarhossain16005@gmail.com (M.M. Hossain), reshmaahmed.sw@gmail.com (R.A. Swarna), rafid.dka@gmail.com (R. Mostafiz), pabonshahacse15@gmail.com (P. Shaha), lubnaju@yahoo.com (L.Y. Pinky), mm73rahman@gmail.com (M.M. Rahman), wahidtuhin0@gmail.com (W. Rahman), selimtee@gmail.com (M.S. Hossain), elias.hossain191@gmail.com (M.E. Hossain), sadiq.iqbal@bu.edu.bd (M.S. Iqbal).

https://doi.org/10.1016/j.mlwa.2022.100330
Received 22 November 2021; Received in revised form 29 January 2022; Accepted 10 May 2022
Available online 17 May 2022

Table 1
Summary of existing works.

Methods	Preprocessing	Feature optimization	Classifier	Accuracy
Polat et al.	–	Best-first search based filter method	SVM	98.5%
Ghosh et al.	Missing values filling	Dimensionality reduction	Gradient boosting	99.80%
Chittora et al.	Rescaling, SMOTE	–	SVM	98.86%
Deepika et al.	–	–	KNN	97%
Gunarathne et al.	Data analysis and rescaling	–	Multiclass decision forest	99.1%
Drall et al.	Data analysis, missing values filling, rescaling	Correlation and dependence approach	KNN	100%
Sharma et al.	Missing values filling, rescaling	–	Decision tree	98.6%

engine provided the maximal accuracy of 98.5% for CKD prediction. Ghosh et al. (2020) established a CKD diagnosis approach using the Gradient boosting (GB) classifier with a feature selection technique. The best accuracy of this model was 99.80%. By analyzing SVM, AdaBoost (AB), Linear discriminant analysis (LDA), and GB classifiers with the feature selection technique the researchers established GB as the best method. (Chittora et al., 2021) analyzed several classifiers along with several feature selection approaches for ML-based CKD diagnosis. SMOTE was used in this system for class balancing and it was gained that Linear SVM with SMOTE without any feature reduction provided the maximum accuracy of 98.86% in their method. Moreover, in this research, a Deep neural network (DNN) based model was also developed and they had found the highest accuracy of 99.6% with their DNN model. Deepika, Rao, Rampure, Prajwal, and Gowda (2020) developed a ML-based real-time application for CKD diagnosis. To build the ML model mentioned method used K-nearest neighbors (KNN) and Naive bayes (NB) classifiers and found the best accuracy of 97% for KNN. Gunarathne, Perera, and Kahandawaarachchi (2017) developed a multiclass decision forest (MCDF) based data mining for CKD detection. By analyzing several algorithms MCDF was chosen as the best option with 99.1% highest accuracy. Drall, Drall, Singh, and Naib (2018) established a CKD prediction model by using the correlation and dependence (CD) technique-based feature selection approach with the KNN classifier. The accuracy of the model was 100%. By analyzing KNN and NB with the CD they had developed the KNN based model. Sharma, Sharma, and Sharma (2016) analyzed 12 different classifications for CKD diagnosis and found the Decision tree (DT) classifier as the best option with 98.6% accuracy. Table 1 presents the summary of the investigation section discussed till now.

From the investigation section, it has been observed that most of the ML-based techniques have satisfactory results. However, to gain those satisfactory results, these methods resort to the assistance of several features optimization techniques. From this point, we have conducted this research to observe the effect of various feature optimization techniques for the diagnosis of CKD.

This research ready-up an existing dataset by performing different preprocessing. Based on the analysis of the ready-up dataset we have observed that an appropriate feature optimization technique may increase the efficiency of the ML-based CKD diagnosis model extensively. From this motive, we have conducted various feature optimization techniques (belong to three different categories) to the dataset and then each optimized dataset is evaluated by an ensemble classifier. The ensemble classifier of this research is made up by using five existing individual classifiers. Leading contributions of this research include:

- Comparative performance analysis of several feature optimization techniques for building the ML-based CKD diagnosis system.
- Finding an automatic CKD diagnosis system with a low false diagnosis rate.
- For CKD diagnosis, Increasing the efficiency of several existing classifiers by making an ensemble model along with different optimization techniques.

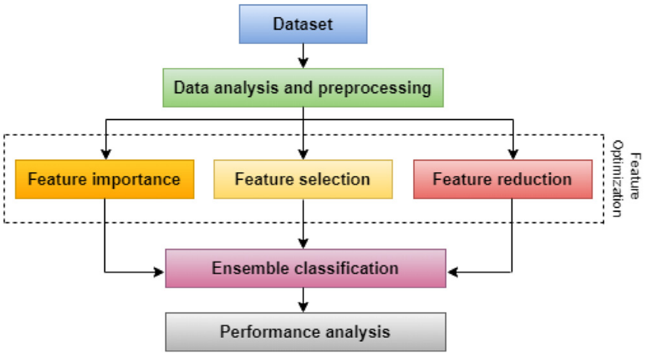


Fig. 1. The architecture of this research.

The next sections of this paper are arranged in the way:- section two presents the related materials as well as the mechanism of the core method of this research; section three presents the results along with the related discussion for this research and finally, section four holds the conclusion of this research.

2. Materials and methodology

The fundamental aim of this research is to identify CKD-infected patients with maximum accuracy by utilizing the machine learning-based approach. To carry out the purpose this research evaluates the effect of several feature optimization techniques to observe their effect and to elect the chief technique. Fig. 1 presents the core architecture of this research at a glance and Sections 2.1 to 2.4 describe the research elaborately.

2.1. Dataset

In this research, the dataset we have used is retrieved from the source Dua and Graff (2019). The dataset consists of 400 chronic kidney disease patients' information having 25 different attributes among them 24 are predictive variables and 1 attribute is decision class. Numeric and nominal are both types of data used for the research. The dataset is clearly described in Table 2.

2.2. Data analysis and preprocessing

There exist two types of data namely nominal and numerical within 24 predicted characteristics of the dataset. For preprocessing this research uses two techniques—filling missing values and converting the nominal values (characteristics that hold binary data) to integer values (i.e- Presenting Normal, and Abnormal by using 0 and 1 that is Normal=0 and Abnormal=1). For missing values, we have utilized the median value to fill the values. After completing preprocessing, we have performed the analysis of the dataset to observe its peculiarity by using

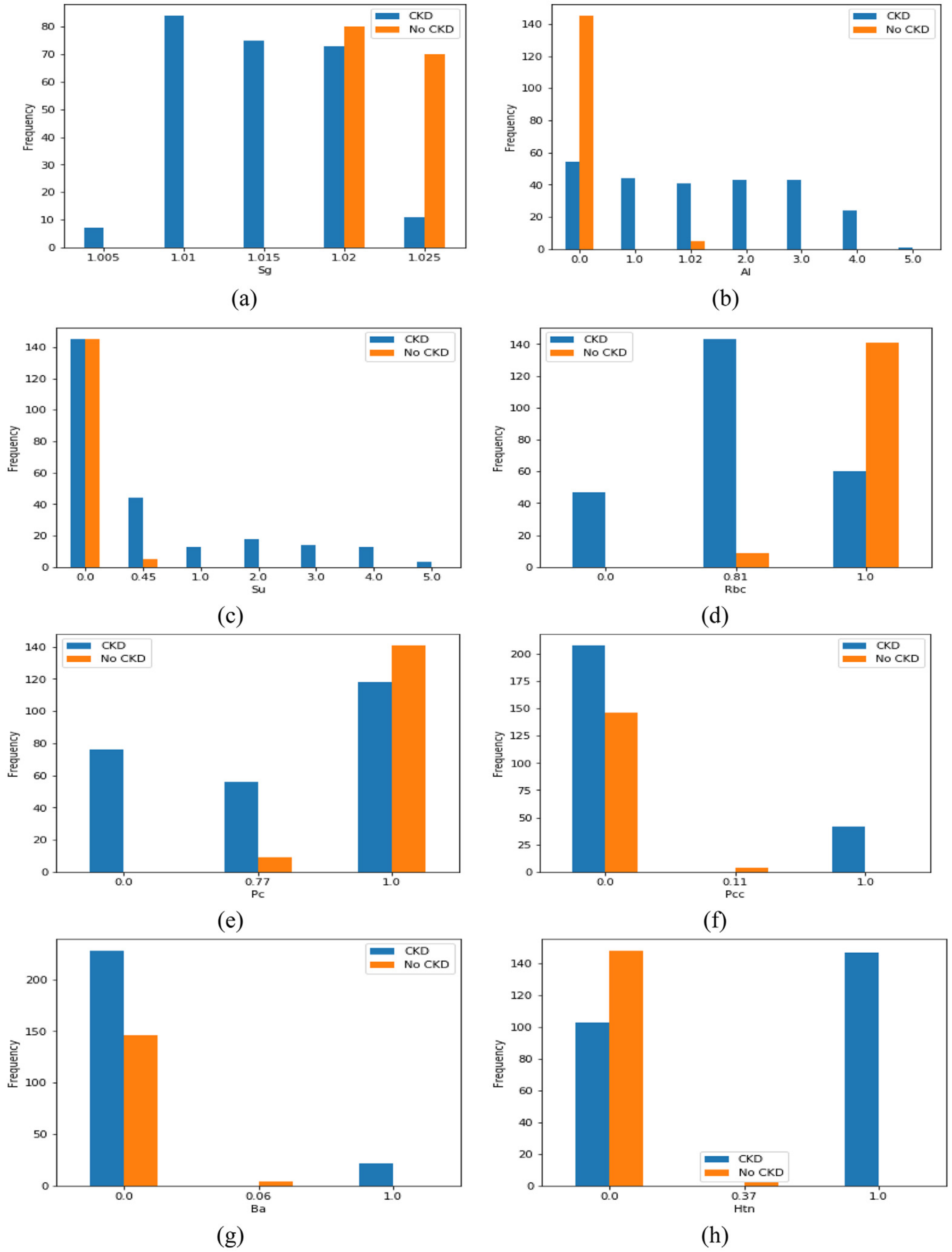


Fig. 2. Characteristics of nominal features: (a) Sg, (b) Al, (c) Su, (d) Rbc, (e) Pc, (f) Pcc, (g) Ba, (h) Htn, (i) Dm, (j) Cad, (k) Appet, (l) Pe, (m) Ane.

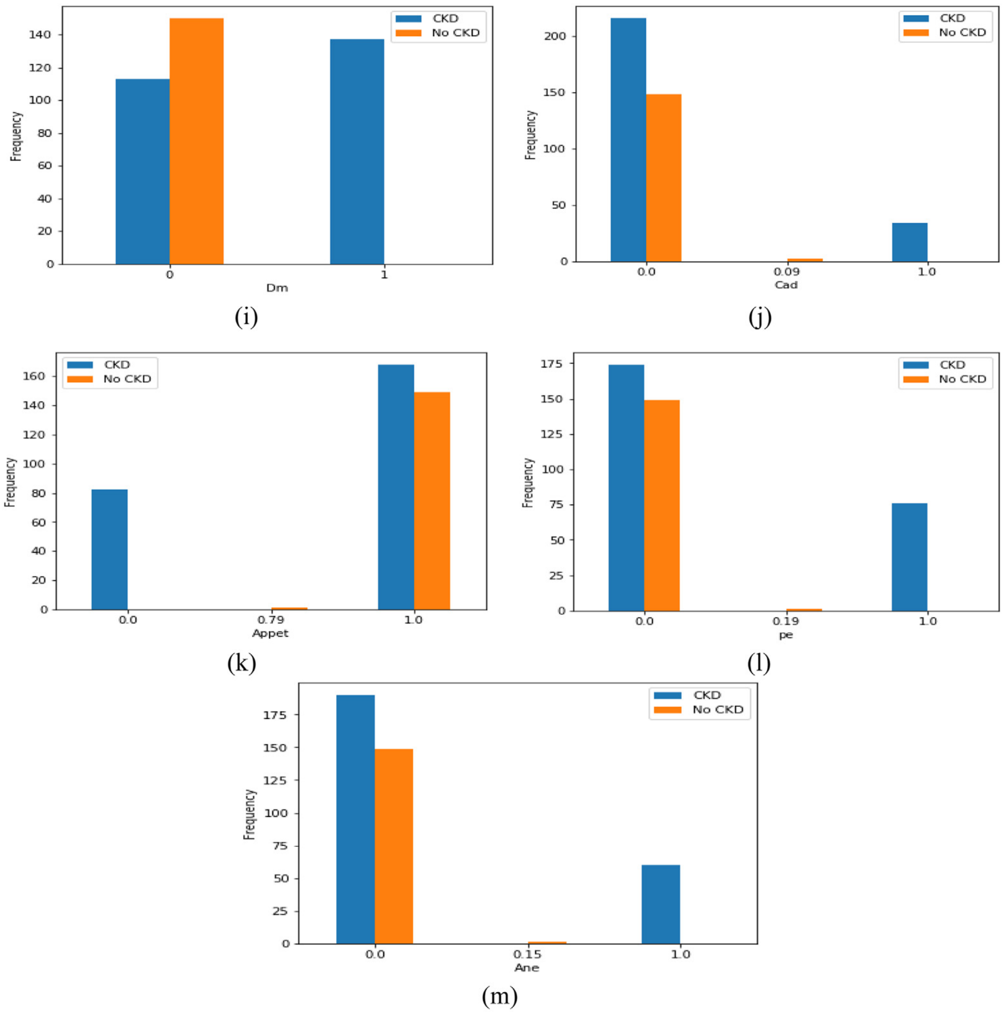
the bar chart and box plot for nominal and numerical characteristics respectively. The characteristics of nominal and numerical features are presented in Figs. 2 and 3 respectively. From Figs. 2 and 3 it has been seen that there exist several outliers and noisy data within the features of the dataset. So, under the above analysis, it can be said that feature optimization may make a great influence to create a highly accurate CKD diagnosis system using an ML-based approach for the mentioned dataset.

2.3. Feature optimization

Feature optimization is the process to minimize the number of features to decrease the complexity of computation as well as to improve the efficiency of the ML model. There exist several approaches for feature optimization and among them, the most common three approaches namely—feature importance, feature selection, and feature reduction are utilized in this research. For each feature optimization approach, we have chosen the two most common and important techniques. Thus, to analyze the strength of feature optimization in ML-based CKD diagnosis this research utilizes six feature optimization techniques. The below subsections describe in detail how each feature optimization approach is utilized in this research.

2.3.1. Feature importance

Feature importance indicates the importance of each feature within a dataset by assigning a certain score value. Based on the score of feature importance we can capture the most important features. Thus,



(i) (j) (k) (l) (m)

Fig. 2. (*continued*).

by eliminating the features containing lower scores we can optimize any feature set. In this research, we have used Decision tree (DT) (Myles, Feudale, Liu, Woody, & Brown, 2004) and Extreme gradient boosting (XGB) (Chen et al., 2015) feature importance techniques.

In DT feature importance technique scores of significant features are found on the minimization of the criterion that is utilized to elect partition points such as entropy or Gini. The DT model gives the property of feature importance after the model gets fit and this property is utilized to capture the importance scores for each input feature.

In the XGB model after constructing boosted trees, the score of importance is measured for each single decision tree through the volume that every attribute partition point enhances the performance quality, weighted through the No. of espials the vertex is responsible for. Finally, by averaging the scores for all decision trees the feature importance is gained.

DT and XGB techniques are chosen since they provide the minimum number of important features for the dataset of this research. Fig. 4 presents the importance of features based on DT and XGB techniques.

2.3.2. Feature selection

Feature selection reduces a feature set by selecting the most reasonable features and by excluding the drossy feature. Two spanking feature selection techniques namely—Minimum redundancy maximum relevance (mRMR) (Radovic, Ghalwash, Filipovic, & Obradovic, 2017) and Recursive feature elimination (RFE) (Yan & Zhang, 2015) are utilized in this research.

mRMR perform feature selection by using the conditions called minimal redundancy (min_D) and maximal relevance (max_R).

$$min_D = \frac{1}{|f|^2} \sum_{i,j \in f} m(i,j) \tag{1}$$

$$max_R = \frac{1}{|f|} \sum_{i \in f} m(k,i) \tag{2}$$

Here f is the set of features and $m(x,y)$ is called Mutual information

$$m(x,y) = \iint q(x,y) \, log \frac{q(x,y)}{q(x)q(y)} dx dy \tag{3}$$

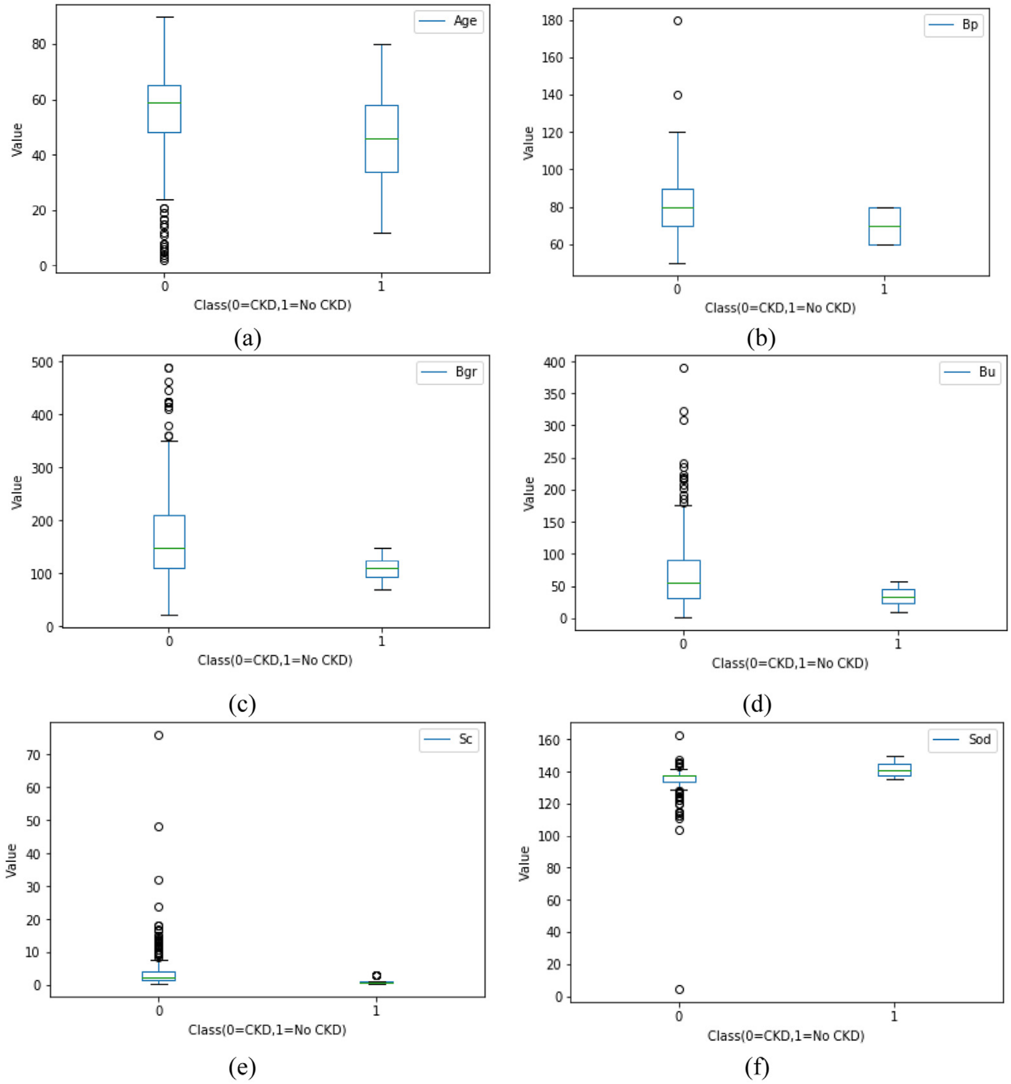
Consider Eqs. (1) and (2) to generate Eq. (4). RFE captures the excellent subset of optimal features by utilizing Eq. (4).

$$maxQ(D,R); Q = D - R \tag{4}$$

By analyzing a number of features we have reserved 8 features for each feature selection technique for the purpose of the best outcome. Table 3 presents the list of selected features for mRMR and RFE.

2.3.3. Feature reduction

Feature reduction reduces a feature set by converting data from higher dimensional space to lower-dimensional space. Two leading and potential feature importance techniques are Principal component analysis (PCA) (Jolliffe, 2005) and Linear discriminant analysis (LDA) (Tharwat, Gaber, Ibrahim, & Hassanien, 2017). So, this research utilizes PCA and LAD for the purpose of feature reduction.



(a) (b) (c) (d) (e) (f)

Fig. 3. Characteristics of numerical features: (a) Age, (b) Bp, (c) Bgr, (d) Bu, (e) Sc, (f) Sod, (g) Pot, (h) Hemo, (i) Pcv, (j) Wbcc, (k) Rbcc.

To do dimensionality reduction PCA generates new uncorrelated features through merging correlated features that gradually maximize variance. To substitute uncorrelated features PCA finds covariance of features through given equation

$$\mathrm{covariance}(X, Y) = \frac{\sum_{j-1}^{m}\left(X_j - \overline{X}\right)\left(Y_j - \overline{Y}\right)}{(m-1)} \tag{5}$$

LDA does dimensionality reduction by conserving class discriminatory information as much as possible. It seeks to capture directions along which the classes may get the greatest dividable by considering the scatter within classes but also the scatter between classes.

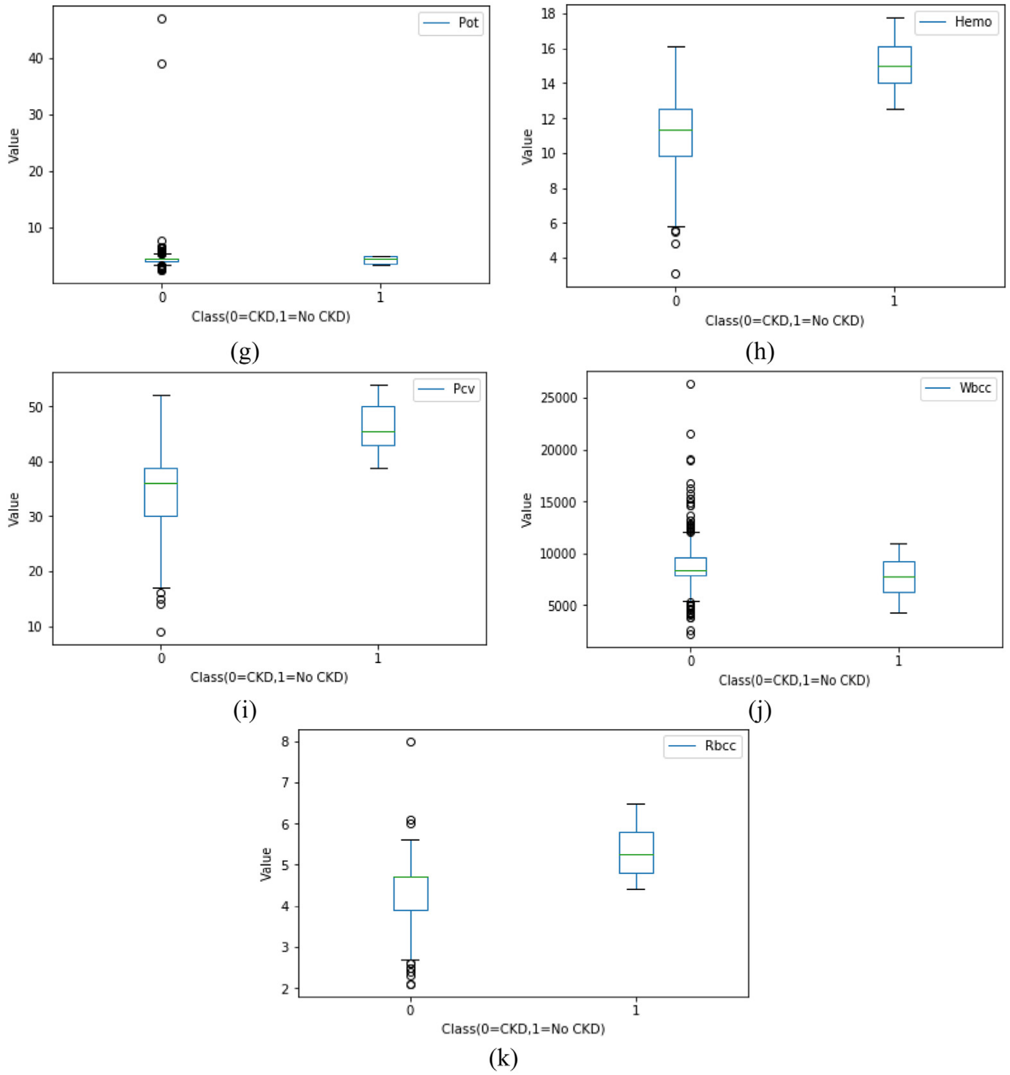
Both PCA and LDA reserve components according to variance ratio. We have reserved 5 features (components) for PCA in this research. Based on the highest variance ratio these 5 components are selected. Fig. 5 presents the variance ratio of various PCA components for the dataset of this research. After applying LDA for feature reduction in the dataset of this research there exist only one component according to the maximum variance ratio. Hence, For LDA 1 feature (component) is reserved in this research.

2.4. Ensemble classification

Each optimized feature set derived from several feature optimization techniques is evaluated individually by an ensemble classifier in this research. Ensemble classifier enhances the results of ML by merging several classifiers. Ensemble process allows generating excellent prediction performance comparing single model (Ren, Zhang, & Suganthan, 2016). This research uses the max voting ensemble approach for CKD prediction. The idea of max voting is to select a certain prediction based on the max vote of a set of classifiers. Five influential classification models namely—Logistic regression (LR) (Dreiseitl & Ohno-Machado, 2002), Random forest (RF) (Liaw & Wiener, 2002), Support vector machine (SVM) (Jakkula, 2006), K-nearest neighbors (KNN) (Laaksonen & Oja, 1996) and Xtreme gradient boosting (XGB) (Chen et al., 2015) are utilized in this research to construct the max voting ensemble model. Fig. 6 depicts the overall architecture of the ensemble model used in this research.

3. Results and discussion

Table 3 presents the number of features each optimization techniques hold after feature optimization. To analyze the efficiency of feature optimization and to elect the best feature optimizer for CKD diagnosis, each of the six feature sets of Table 4 is examined individually by mentioned max voting ensemble model. During execution by ensemble model, each feature set is partitioned into two portions 80% data for training and 20% for testing purposes. The whole process is evaluated by using 10-fold cross-validation.



(g) (h)

(i) (j)

(k)

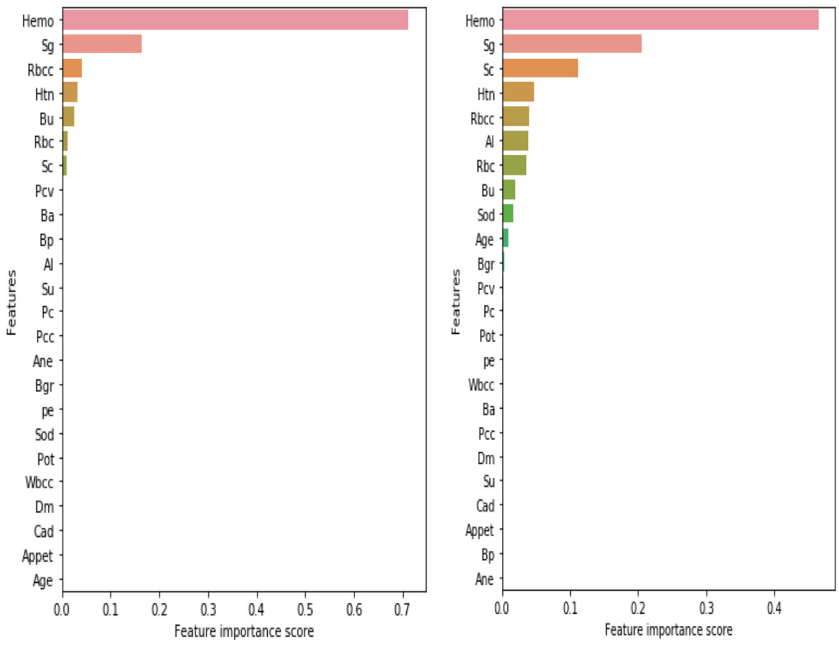
Fig. 3. (*continued*).

Confusion matrix (CM) shows the actual and miss detection of any ML model in a tabular form. Fig. 7 illustrates the formation of CM for this scheme. Fig. 8 shows the CM in normalized form (presenting actual and miss detection in percentage form) for different techniques of this research. Figs. 9 and 10 summarize the outcome of Fig. 8 at a glance. Where Fig. 9 presents the overall miss detection rate for CKD positive cases, using different techniques and Fig. 10 presents the overall miss detection rate for CKD negative cases, using different techniques. Figs. 9 and 10 prove that the feature optimization technique LDA outperforms all other techniques.

Six performance measurement parameters namely—accuracy, specificity, sensitivity, precision, F1-score, and Kappa statistic (kappa) are used in this research to evaluate the performance of the ensemble model. Table 5 presents the description of these performance measurement metrics in detail. Table 6 shows the values of performance measurement metrics for various methods of this research. Fig. 11 presents the overall miss detection rate of different techniques of this research. The miss detection rate of Fig. 11 shows the related inaccuracy of each feature optimization approach on the dataset of this research. For instance, Fig. 11 presents that LDA contains an overall miss detection rate of 0.5 which indicates feature optimization technique LDA provides 100-0.5 =99.5% accuracy for our scheme. Thus Fig. 11 presents nothing but the comparison of the overall accuracy of different feature optimization approaches for this research. By analyzing Table 6 and Fig. 11 it is observed that LDA with ensemble model provides the best approach for diagnosis of CKD.

The receiver operator characteristic (ROC) curve is considered as the binary classification issue evaluation matric that is used to illustrate the probability curve. This curve shows the rate of TP against the rate of FP at the different threshold levels. The Area Under the Curve (AUC) is the summary of the ROC curve that is used for measuring the classifier's ability to distinguish between classes. The AUC shows how well the model is distinguished between the positive and negative classes. While the AUC is increasing, the model will get better. When the AUC=1, the classifier is distinguishable between all positive and negative class points. If the AUC becomes 0, then the classifiers will expect all the negatives to be positives and all the positives to be negatives. When AUC=0.5, the classifier cannot differentiate between the positive and the negative class points. That means the classifier is only capable of predicting either a random class or a constant class for all data points. As a result, the ability to distinguish between the positive and negative classes is increased with the classifier AUC's higher value (Bradley, 1997). For different methods of this research, Fig. 12 shows the ROC-AUC curve of each fold and the overall system.

For both mRMr and RFE, the peak outcome is gained with a minimum of 8 features hence 8 features are reserved for these methods.



(a) (b)

Fig. 4. Score of each feature based on: (a) DT feature importance; (b) XGB feature importance.

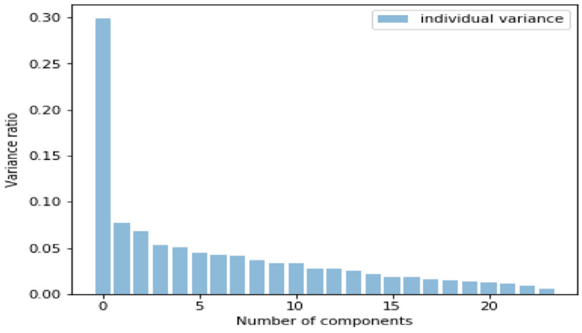
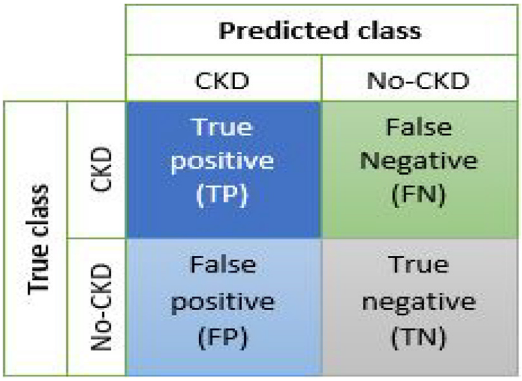


Fig. 5. Variance ratio of various PCA components.



		Predicted class	
		CKD	No-CKD
True class	CKD	True positive (TP)	False Negative (FN)
	No-CKD	False positive (FP)	True negative (TN)

Fig. 7. Formation of CM for this scheme.

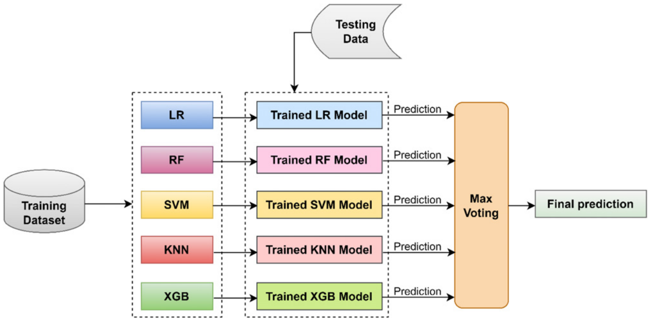


Fig. 6. The architecture of max voting ensemble model used in this research.

Fig. 13 shows the overall accuracy of mRMr and RFE by using the different number of features. In Fig. 13, the x-axis presents the number of features and the y-axis presents the corresponding accuracy. Fig. 13 shows that for the number of features 8 both mRMr and RFE hold their maximum accuracy and with the number of features more than 8 either they hold the accuracy as same as with 8 number of features or less than of that.

For PCA the peak outcome is gained with a minimum of 5 components hence 5 components are reserved for this method. Fig. 14 shows the overall accuracy of PCA by using the different number of components. In Fig. 14, the x-axis presents the number of components and the y-axis presents the corresponding accuracy. Fig. 14 shows that for the number of components 5 PCA holds its maximum accuracy and with the number of components more than 5 either it holds the accuracy as same as with 5 components or less than of that.

For LDA-based dimensionality reduction, this research uses 1 component. The dataset of this research returns only one component after applying LDA to it. As there exists only 1 component hence 1 component is reserved for LDA and Fig. 15 shows the visualization of this component with respect to each class (0=CKD, 1=No CKD). Fig. 15 shows that the single LDA component provides clear separability between the two classes.

Table 7 shows the comparison between the proposed method and the existing methods. For comparison, we have considered the works

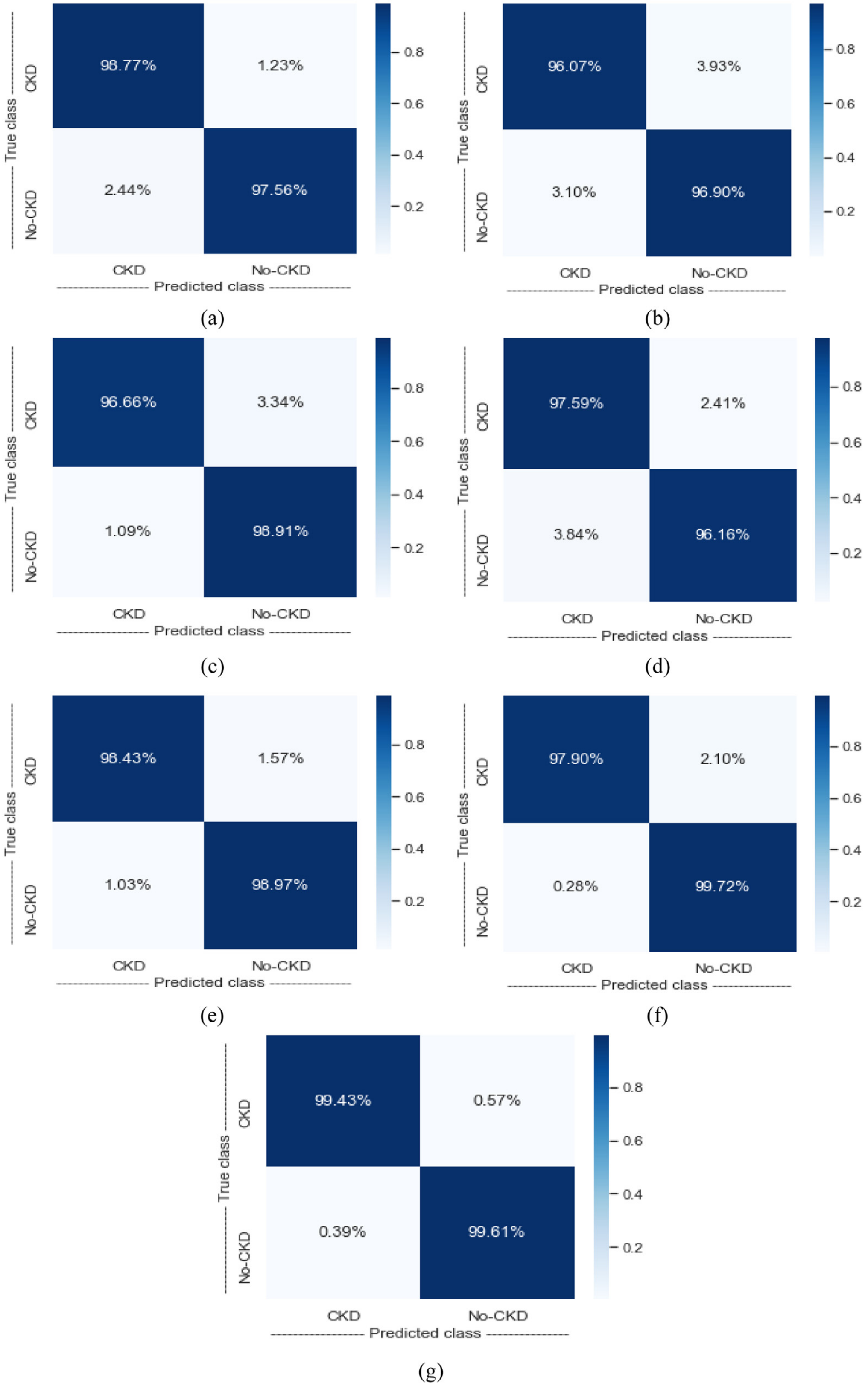


Fig. 8. Confusion matrix of the system: (a) Using original features, (b) Using decision tree feature importance, (c) Using XGB feature importance, (d) using mRMr feature selection, (e) Using RFE feature selection, (f) Using PCA feature reduction and (g) Using LDA feature reduction.

that utilize the same dataset as us. From Table 7 it has been seen that some methods hold 100% accuracy but considering our method we can overlook this performance since most of them do not use any fold cross-validation. Moreover, if we look at Fig. 12, we can find that in several ROC curves for a particular fold the AUC value is 1 and this indicates the overall accuracy of this research is 100% for that fold.

Table 2
Presentation of the CKD dataset used in this research.

Feature name	Description	Type	Values
Age	Age of the patient	Numeric	Years
Bp	Blood Pressure	Numeric	mm/Hg
Sp	Specific Gravity	Nominal	(1.005, 1.010, 1.015, 1.020, 1.025)
Al	Albumin	Nominal	(0, 1, 2, 3, 4, 5)
Su	Sugar	Nominal	(0, 1, 2, 3, 4, 5)
Rbc	Red Blood Cells	Nominal	(Normal, Abnormal)
Pc	Pus Cell	Nominal	(Normal, Abnormal)
Pcc	Pus Cell Clumps	Nominal	(Present, Not Present)
Ba	Bacteria	Nominal	(Present, Not Present)
Bgr	Blood Glucose Random	Numeric	mgs/dl
Bu	Blood Urea	Numeric	mgs/dl
Sc	Serum Creatinine	Numeric	mgs/dl
Sod	Sodium	Numeric	mEq/L
Pot	Potassium	Numeric	mEq/L
Hemo	Hemoglobin	Numeric	Gms
Pcv	Packed Cell Volume	Numeric	Percentage
Wc	White Blood Cell Count	Numeric	cells/cumm
Rc	Red Blood Cell Count	Numeric	millions/cmm
Htn	Hypertension	Nominal	(Yes, No)
Dm	Diabetes Mellitus	Nominal	(Yes, No)
Cad	Coronary Artery Disease	Nominal	(Yes, No)
Appet	Appetite	Nominal	(Good, Poor)
Pe	Pedal Edema	Nominal	(Yes, No)
Ane	Anemia	Nominal	(Yes, No)
Class	Decision Class	Nominal	(ckd, notckd)

Table 3
Features selection using mRMR and RFE techniques.

Technique	Selected features
mRMr	Bgr, Wbcc, Bu, Pcv, Bp, Sod, Sc, Hemo
RFE	Al, Su, Rbc, Hemo, Rbcc, Htn, Dm, Pe

Table 4
The number of features for various methods of this research.

Method	Technique	No of features
Feature importance	DT	7
	XGB	10
Feature selection	mRMr	8
	RFE	8
Feature reduction	PCA	5
	LDA	1

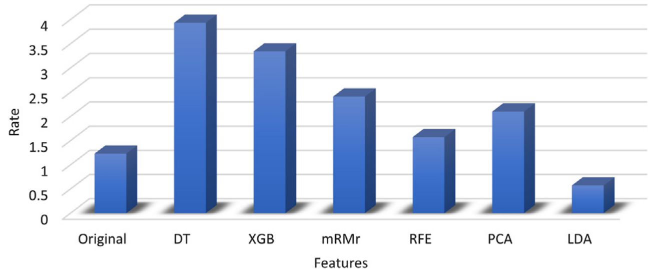


Fig. 9. Overall miss detection rate for CKD positive cases, using different techniques.

Table 5
Description of performance measurement metrics.

Metrics	Formula	Meaning
Accuracy	$\frac{TP+TN}{FP+TP+FN+TN} \times 100$	Rate of overall correct prediction.
Specificity	$\frac{TN}{TN+FP} \times 100$	Rate of correct No-CKD prediction cases
Sensitivity	$\frac{TP}{TP+FN} \times 100$	Rate of correct CKD prediction cases.
Precision	$\frac{TP}{TP+FP} \times 100$	Rate of correct CKD prediction cases from all predicated CKD cases.
F1-score	$2 \times \frac{\text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}} \times 100$	Describe how precise and robust a classifier is
Kappa	$\frac{\text{Accuracy} - \text{Expected accuracy}}{1 - \text{Expected accuracy}}$	Indicate the efficiency of a classifier. The more the value of kappa closer to 1 the better the classifier is.

Table 6
Performance of measurements for various techniques.

Features	accuracy	Specificity	Sensitivity	Precision	F1 score	kappa
Original	98.38	97.56	98.77	98.65	98.70	0.965
DT	96.38	96.90	96.07	98.19	97.11	0.923
XGB	97.50	98.91	96.66	99.41	98.00	0.946
mRMr	97.13	96.16	97.59	97.84	97.71	0.938
RFE	98.63	98.97	98.43	99.41	98.91	0.970
PCA	98.5	99.72	97.90	99.78	98.82	0.967
LDA	99.5	99.61	99.43	99.82	99.62	0.989

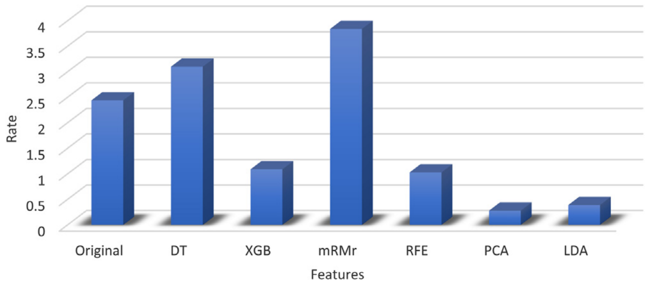


Fig. 10. Overall miss detection rate for CKD negative cases, using different techniques.

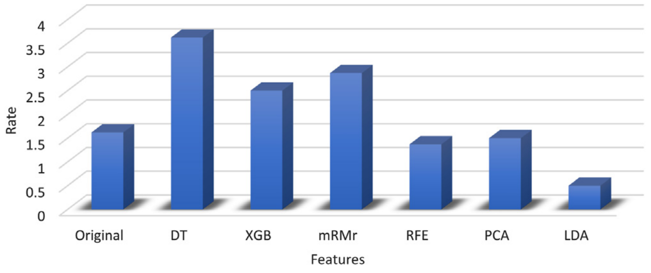
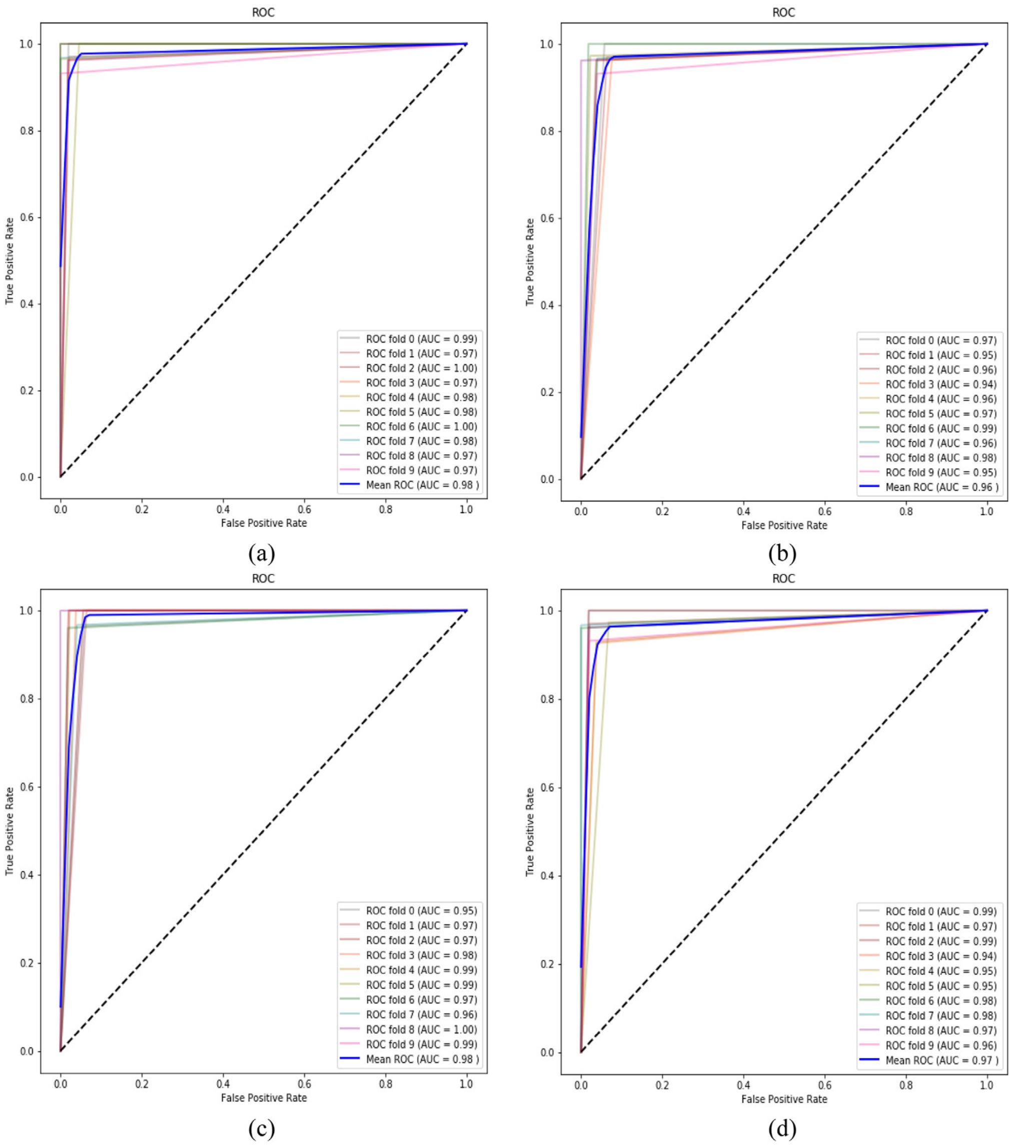


Fig. 11. Overall miss detection rate using different techniques.

4. Conclusion

Early detection of CKD is an essential step to getting proper treatment. Hence this research builds a high accurate automatic CKD detection model from various clinical attributes of a patient. The clinical attributes are considered as features in this scheme and the model is built by using the ML approach along with feature optimization techniques. There exist different feature optimization approaches. Although it is assumable that feature optimization can increase the efficiency of any ML model but for this purpose, the right feature optimization technique needs to choose. Hence this research analysis several feature optimization techniques along with the ML model for diagnosis CKD to see how they affect the performance of the ML model. Based on our analysis we have found that not all feature optimization techniques perform well along with our ML model. In some cases, the feature optimizer performs poorly than the ML model performs without any



(a)

(b)

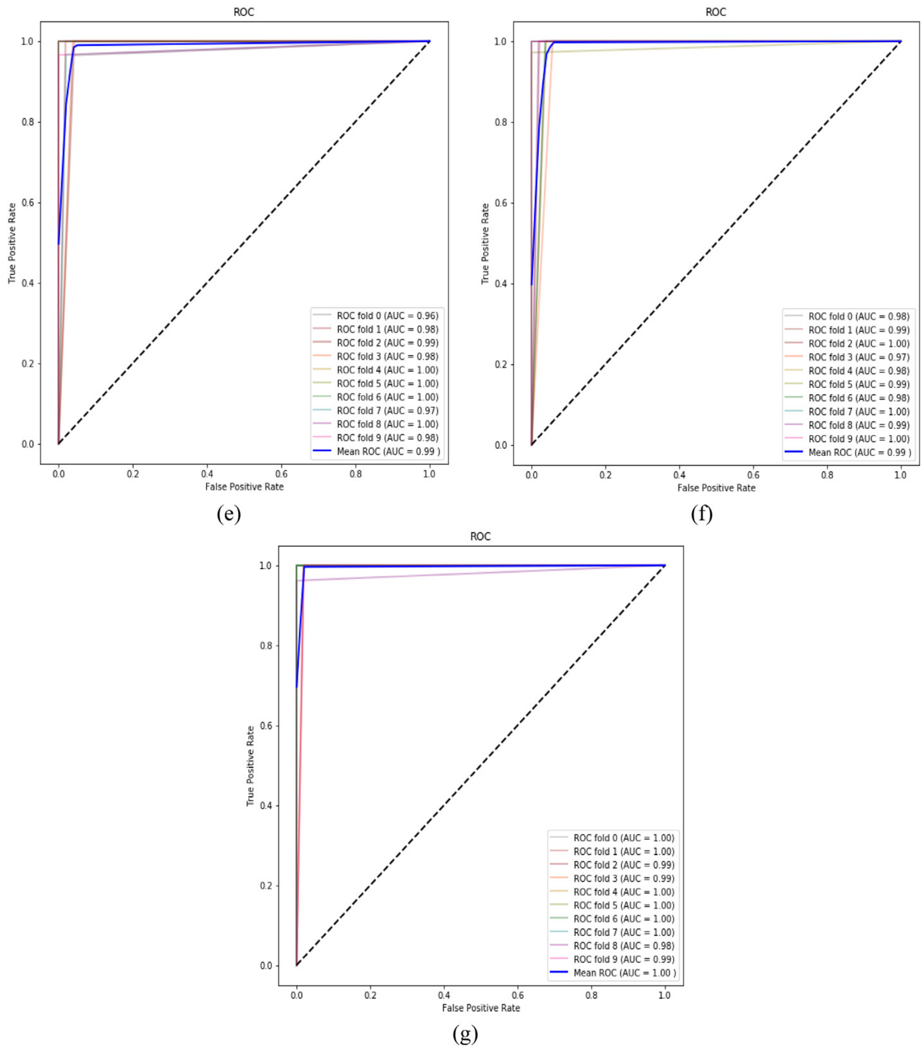
(c)

(d)

Fig. 12. ROC curve of the system: (a) Using original features, (b) Using decision tree feature importance, (c) Using XGB feature importance, (d) using mRMr feature selection, (e) Using RFE feature selection, (f) Using PCA feature reduction and (g) Using LDA feature reduction.

Table 7
Comparison of different ML-based CKD diagnosis systems.

Methods	Technique	Accuracy
Polat et al	Filter feature selection + SVM	98.5 (using 10-fold)
Ghosh et al.	Feature selection + GB	99.80 (using no fold)
Chittora et al.	SMOTE + Linear SVM	98.86 (using no fold)
Deepika et al.	KNN	97 (using no fold)
Drall et al.	CD feature selection + KNN	100 (using no fold)
Proposed method	LDA feature reduction + Max voting Ensemble classification	99.5% (using 10-fold) 100% (using no fold)
Performance of LDA with different classifier	LDA + SVM	98.5 (using 10-fold) 98.75 (using no fold)
	LDA + GB	98.125(using 10-fold) 100 (using no fold)
	LDA + Linear SVM	98.25 (using 10-fold) 98.75 (using no fold)
	LDA + KNN	97.875 (using 10-fold) 98.75 (using no fold)



(e)

(f)

(g)

Fig. 12. (*continued*).

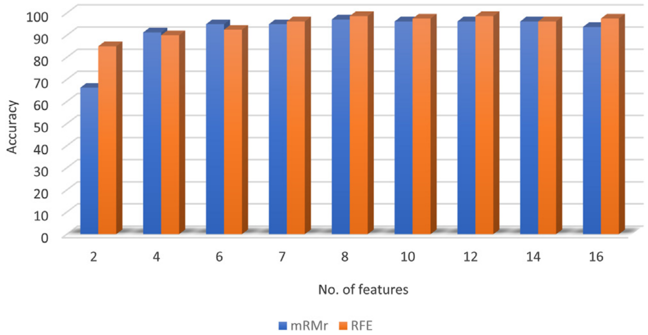


Fig. 13. Accuracy of feature selection methods (mRMr and RFE) with respect to different number of features.

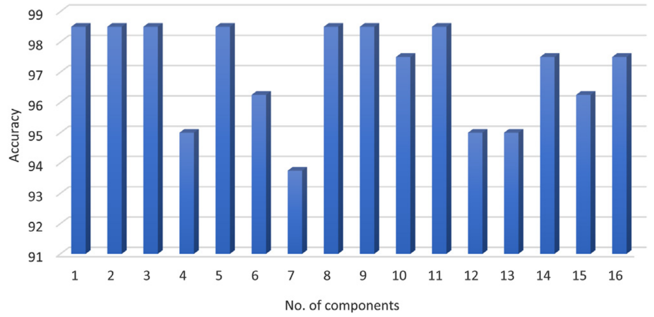


Fig. 14. Accuracy of feature reduction method (PCA) with respect to different number of components.

feature optimization. However, based on our analysis we have found that feature optimizer LDA performs highest with our ML model by giving an overall inaccuracy of 0.5%, where the max inaccuracy of the ML model without any feature optimization is 1.62%. Due to the lack dataset, this research utilizes all actions on a single UCI dataset so this can be a limitation of our work. Since the decision based on experimenting with a single dataset may not be adequate. In the future, we will try to work with more than one dataset. We will also expand the work by using more feature optimization techniques along with different ML approaches.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

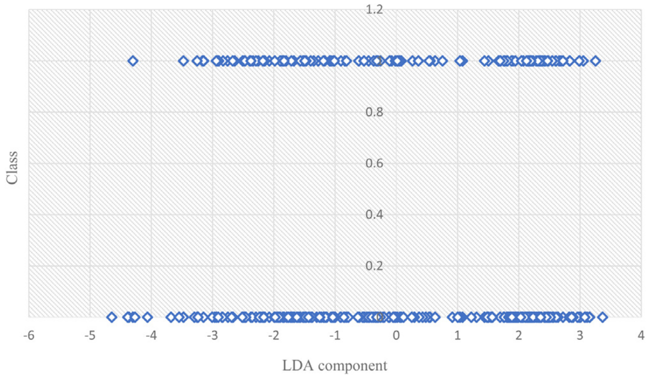


Fig. 15. Visualization of LDA component for each class.

References

Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications*, *10*(8).

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1–4.

Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z. . . ., et al. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access*, *9*, 17312–17334.

Chronic Kidney Disease in the United States, 2021 (2021). Retrieved from centers for disease control and prevention. https://www.cdc.gov/kidneydisease/publications-resources/ckd-national-facts.html.

Deepika, B., Rao, V. K. R., Rampure, D. N., Prajwal, P., & Gowda, D. G. (2020). Early prediction of chronic kidney disease by using machine learning techniques. *American Journal of Computer Science and Engineering Survey*, *8*(2), 7.

Drall, S., Drall, G. S., Singh, S., & Naib, B. B. (2018). Chronic kidney disease prediction using machine learning: A new approach. *International Journal of Management, Technology and Engineering*, *8*(5), 278–287.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359.

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science, [http://archive.ics.uci.edu/ml].

Ghosh, P., Shamrat, F. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020). Optimization of prediction method of chronic kidney disease using machine learning algorithm. In *2020 15th International joint symposium on artificial intelligence and natural language processing* (pp. 1–6). IEEE.

Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In *2017 IEEE 17th International conference on bioinformatics and bioengineering* (pp. 291–296). IEEE.

Jakkula, V. (2006). *Tutorial on support vector machine (Svm)* (p. 37). School of EECS, Washington State University.

Jolliffe, I. (2005). Principal component analysis. In *Encyclopedia of statistics in behavioral science*.

Laaksonen, J., & Oja, E. (1996). Classification with learning k-nearest neighbors. In *Proceedings of international conference on neural networks, Vol. 3* (pp. 1480–1483). IEEE.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), 275–285.

Nikhila (2021). Chronic kidney disease prediction using machine learning ensemble algorithm. In *International conference on computing, communication, and intelligent systems* (pp. 19–20).

Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of Medical Systems*, *41*(4), 55.

Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, *18*(1), 1–14.

Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, *11*(1), 41–53.

Sharma, S., Sharma, V., & Sharma, A. (2016). Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. arXiv preprint arXiv:1606.09581.

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, *30*(2), 169–190.

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B (Chemical)*, *212*, 353–363.